MANAV RACHNA INTERNATIONAL UNIVERSITY

Short Synopsis

For

Ph. D. Programme 2013-14

Title: Design and Analysis of Techniques for Clustering of Text Documents

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

FACULTY OF ENGINEERING & TECHNOLOGY

Submitted by:

Name: Maedeh Afzali
Signature:
Registration No.: 13/Ph.D./034
Email ID: maedeh.af@gmail.com
Contact No.: +91-8860270443

Supervisor:

Name: Dr. Suresh Kumar
Signature:
Designation: Professor
Email ID: suresh.fet@mriu.edu.in
Contact No.: +91-9810487182

FACULTY OF ENGINEERING & TECHNOLOGY

Declaration

This is to declare that this short synopsis reflects the research work design and its contents are the original piece of work, as per guidelines, grammatically correct and free from plagiarism. This has been prepared under the guidance and supervision of my supervisor.

Signature of Research Scholar
Name : Maedeh Afzali
Registration No. : 13/Ph.D./034
Batch : 2013-2014
Department : Computer Science & Engineering

Certificate

This is to testify that the entire text of this Synopsis, prepared by my research scholar, has been read thoroughly and it is as per the guidelines and grammatically correct. Further, it is original piece of her work and free from plagiarism.

Signature of the supervisor
Name : Dr. Suresh Kumar
Designation : Professor
Department : Computer Science & Engineering

Recommendation

I recommend this short synopsis prepared and submitted by **Ms. Maedeh Afzali** for approval of Ph.D. Quality Enforcement Committee and thereafter to be uploaded on Shodh Gangotri.

(Dr. M K Soni)
Executive Director & Dean, FET, MRIU

ABSTRACT

Nowadays, we are in an era where tremendous amount of data is being generated through various sources, resulting in emergence of a new kind of heterogeneous, fast and high-volume data, which is called Big Data. Analyzing and processing these data through existing data mining techniques is burdensome. Therefore, it is required to develop innovative mining techniques based on new frameworks and platforms that have the potential to successfully overcome these challenges. Moreover, majority of data that is produced today is in text format, which require hardship to analyze and process. The main focus in this research is designing efficient clustering techniques in Hadoop and MapReduce framework to cluster the text documents which will be helpful in information retrieval, browsing and searching, etc. In document clustering selecting an appropriate similarity measure to find the closeness between the documents is a challenging task.

The author will focus on using clustering techniques, to cluster the text documents. Prior to clustering an investigation is required to select an appropriate similarity measure to find the similar documents to achieve coherent clusters. A new similarity measure will be proposed in order to develop and design an efficient algorithm for text document clustering.

***Key words**: Hadoop, MapReduce, Data Mining, Clustering, Document Clustering, Similarity Measures*

CONTENTS

S. No.	Description	Page No.
1	Introduction	1-2
2	Literature Review	2-4
	2.1 Big Data Description	2-3
	2.2 Finding Appropriate Similarity Measure	3
	2.3 Clustering using Hadoop and MapReduce	3-4
	Description of Broad Area/Topic	4-13
	3.1 Characteristics of Big Data	4-6
	3.2 Hadoop	6
	3.2.1 Hadoop Architecture	6-7
	3.2.2 Hadoop Distributed File System (HDFS)	7
	3.3.3 MapReduce	8
	3.3 Data Mining	9
	3.3.1 Clustering	9
	3.3.2 K-Means	9
	3.3.3 Agglomerative Hierarchical Clustering	10
	3.4 Document Representation	10
	3.5 Similarity Measures	11-13
4	Challenges	13-14
5	Problem Description	14
6	Objective of Research	14-15
7	Methodology	15
8	Expected Outcome of the Research	15
9	Proposed Time Frame	16
10	References	17-19

1. INTRODUCTION

Nowadays, we are in an era where tremendous amount of data is being generated through various sources such as variety of digital devices, remote sensors, social networking profiles and government processes, which results in emergence of a new science called Big Data. According to O'Reilly, an American Media Company, Big Data is defined as the data that is large enough that it exceeds the processing capabilities of conventional database architectures (White**,** 2012). The size of data is too big; it moves too fast and does not fit in the structure of traditional databases. Unfortunately, analyzing and processing this kind of data through existing data mining techniques require hardship, which sometimes might be frustrating. Therefore, there is a need to develop innovative mining techniques based on new frameworks and platforms that will have the potential to successfully overcome the challenges that is faced in Big Data. Moreover, the majority of data that is produced nowadays is in the text format. Seth Grimes (Breakthrough, 2008) a leading industry analyst, having specialization in data analysis technologies published an article that stated, "Approximately 80% of business-relevant information originates in unstructured form, primarily text." Consequently, there is a requirement of analyzing this kind of data, in order to extract the hidden knowledge and insights from this data.

Data mining is a technology that help us to efficiently discover interesting information, hidden beneath the huge amounts of data. Basically, it turns a large collection of data into useful knowledge. One of the most significant method of data mining is clustering, which is an unsupervised learning technique, where the class label of input data is unknown. It groups the data objects into clusters in such a way that the objects within a cluster are similar to each other and dissimilar to the objects in other clusters. Clustering of text documents is a highly active research area. It can aid the organizations to group their documents to achieve an efficient browsing and searching.

Before, clustering the documents, an appropriate similarity measure has to be determined, which aids us to find the documents which are close to each other, therefore, highly coherent clusters are achieved. In other words, similarity measure is a function that calculates the degree of similarity between the pair of documents. There is verity of

similarity measures available, but it's not clear that which similarity measure is more effective in document clustering. Understanding the effectiveness of these similarity measures and selecting an appropriate measure is a challenging task.

The author in this research, first will try to compare several similarity measures and find out which measure fits best for clustering the documents. Secondly, will investigate that which clustering algorithm provides a better behavior while clustering. Thirdly, will propose a new similarity measure to develop and design an efficient clustering algorithm in order to cluster the text documents.

2. LITERATURE REVIEW

The literature review has been divided into three sections. Section 2.1 provides a review on some papers, which provides a broad overview on Big Data and Hadoop framework. Section 2.2 deals with the attempts that have been done to find appropriate similarity measure for text document clustering. Finally, in section 2.3 the work related to finding the similarity of documents in Hadoop and MapReduce framework is discussed.

2.1 In this section some papers that provide a definition for big data concept and Hadoop framework is discussed:

Nowadays, big data is rapidly growing in all fields of science and engineering. The data collection has increased rapidly, and the existing data mining tools don't have the capability to capture, manage and process these data. Wu et al. (2008), have presented HACE theorem that characterizes the features of Big Data as large-volume, heterogeneous, autonomous data sources, with distributed and decentralized control and try to explore complex relationships among data. Moreover, they have provided a conceptual view of the Big Data processing framework, which is in three tiers, considerations on data accessing and computing (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III).

Dean and Ghemawa (2008), have discussed MapReduce programming model, which is used for providing implementations for processing and analyzing large datasets in Hadoop environment. The user provides the Map and Reduce function, to parallelize the computation across the large cluster of machines to make the analyzing process efficient.

2.2 Attempts that has been done by some authors to find out the appropriate similarity measure for text document clustering is discussed below:

Steinbach et al. (2000), compared two main approaches for clustering the documents, agglomerative hierarchical algorithm and k-means (standard k-means and bisecting k-means). To evaluate the quality of clusters the entropy, F-measure and UPGMA are used. The result of investigation showed that the bisecting k-means technique produce significantly better results as compared to hierarchical approaches. Moreover, they claim that the results of clustering is highly dependent on the nature of documents.

Huang (2008), compared the effectiveness of five similarities measures namely Euclidean distance, Cosine similarity, Jaccard coefficient, Pearson's correlation coefficient and Averaged kullback-Leibler divergence. Then most popular partitional clustering algorithm, k-means algorithm is used to cluster the text documents. Finally, the evaluation was done on some available commonly used text document data sets. The result demonstrated that except Euclidean distance measure, the other measures have considerable effectiveness and produce better clustering results.

El-Hamdouchi (1989), used different hierarchical agglomerative approaches like, single linkage, complete linkage, group averaged and Ward, in order to cluster the text documents. The comparison results illustrate that group averaged and complete linkage method provided a better result as compared to other approaches.

2.3 This section discusses the work that has been done to find the document similarity using Hadoop framework:

Elsayed, Lij and W., Oard (2008) proposed a MapReduce algorithm to compute pairwise document similarity in large collection of documents. The reason of using MapReduce

framework was because they could break down the inner products calculations in computing document similarity into separate multiplication and summation stages in such a way that make their computation efficient through distributing it across several machines.

Niyigena et al. (2015), provided a method to detect near duplicate documents in large corpora. Moreover, they provided a new method to compute pairwise document similarity in corpus to reduce the time execution and save space resources. In their proposed method instead of considering the computation of all pair's candidates which is used in existing algorithms, they have reduced the number of candidate's pairs to be compared, which is highly beneficial in reducing the execution time and space.

3. DESCRIPTION OF BROAD AREA

In today's technical world, the amount of data is drastically increasing, and also the data is being generated in different formats. Based on the report given by IBM, approximately 2.5 quintillion bytes of data is generated every day. Moreover, approximately 90 % of data in the world today has been generated in last two years. This large and complex collection of data is called Big Data.

3.1 Characteristics of Big Data

The following characteristics describe the Big Data.

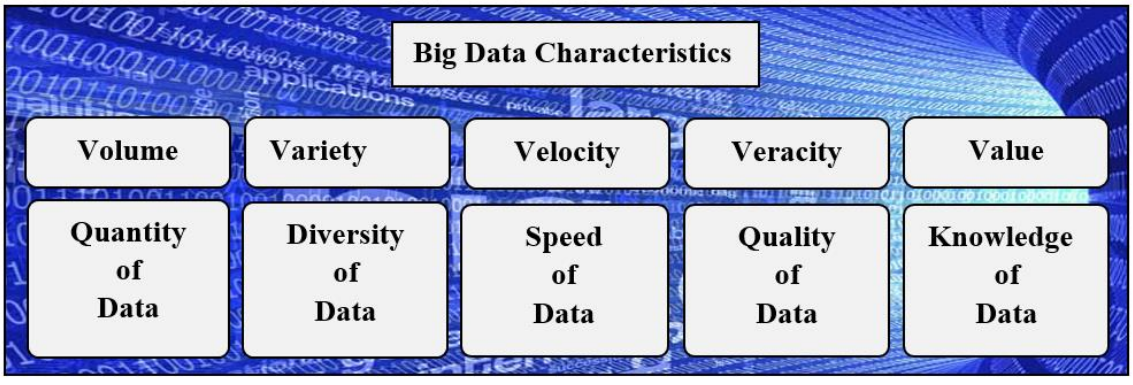

Figure 1: Big Data Characteristics

Volume: The massive amount of data being produced by organizations, outpace traditional storage and analytical solutions. Twitter generates 7 terabytes data every day. Facebook generates 10 terabytes. It is anticipated that the scale of data will be increased from 0.8 million Petabytes in 2000 to 35 Zettabytes in 2020.

Variety: The data that is produced in today's technical world comes from various sources in different formats. The formats are in three types:

- **Structured -** The structured data can be easily stored, organized and analyzed. These data are demonstrated in organized tables consisting of columns and rows.
- **Semi-structured -** The data is organized based on a particular schema and the information associated with it such as metadata, that makes the organizing and analyzing easier.
- **Unstructured –** The heterogeneous data that does not have an identifiable structure. This data comes in various formats such as text, image, video and more.

Velocity: Nowadays, data is generated in real time. Velocity describes that how fast the data is coming and how quickly it is stored and retrieved.

Veracity: When there is availability of the large amount of data, there is a demand to ensure that the data being processed is accurate and complete (Ishwarappa and Anuradha, 2015).

Value: Big data help organizations to generate values from their data. It has the potential to enable the users to get better results from the data stored and characterize what value could come out of which data.

To handle the upcoming challenges because of the characteristics of Big Data numerous attempts have been made over years. The early step was taken by Doug cutting who originated the Apache Nutch (Apache Nutch™, 2004). Thereafter in 2003, Google provided a scalable distributed file system to support increasing demands of Google's data processing needs (Ghemawat, Gobioff, and Leung, 2003) called Google File System (GFS). In 2006, when Doug Cutting was hired by Yahoo, based on GFS concept, he and his dedicated team designed a new framework called Hadoop to handle problems that are too big and are not solvable by regular systems. Moreover, in 2008 Hadoop became Apache Top Level Project. Apache Hadoop is an open-source software that provides distributed

processing of large data sets with the help of simple programming models (Shvachko et al., 2013). Today many organizations such as Yahoo, Facebook, Ebay, Amazon, Twitter, LinkedIn uses Hadoop because of its low cost, scalability, fault tolerance and flexibility.

3.2 Hadoop

As depicted in Figure 2, Hadoop is designed based on two main components called Hadoop Distributed File System (HDFS), a storage file system and MapReduce, a parallel programming framework. In addition to these two components, there are some additional tools, such as Apache Yarn, Apache Pig, Apache Hive, Apache HBase, Apache Spark, Apache Oozie, Apache Mahout, etc., build on top of Hadoop in order to expand the analytical and processing capabilities of Hadoop.

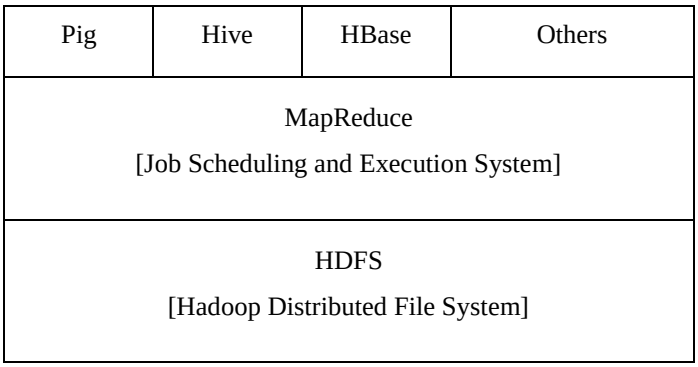

Figure 2: Hadoop Ecosystem

3.2.1 Hadoop Architecture

Hadoop architecture is based on a master-slave model. As shown in Figure 3, there is one NameNode, which operates as master node and several DataNodes, which operates as slave nodes. Basically, there are three daemons functioning in HDFS, namely NameNode, Secondary NameNode and DataNodes and two daemons JobTracker and TaskTracker functioning on MapReduce (Alam and Ahmed, 2014).

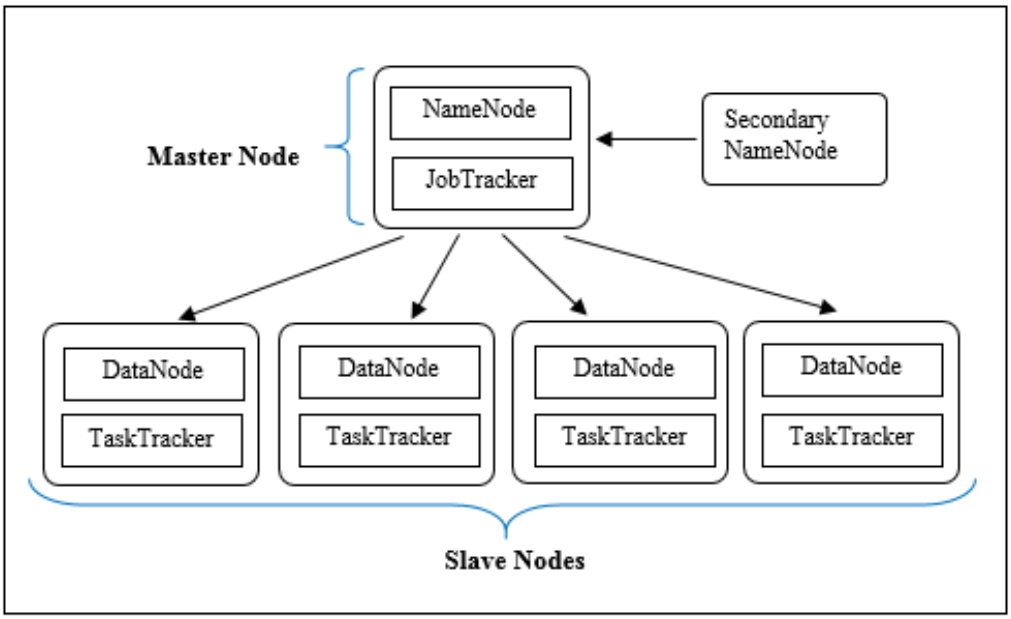

Figure 3: Hadoop Architecture

3.2.2 Hadoop Distributed File System (HDFS)

Hadoop distributed file system (HDFS) is a file storage system, which mainly provides distributed storage of a huge number of information, across a collection of servers within a cluster. Ideally, in HDFS the user's data is partitioned into blocks with default size of 64 MB. As mentioned earlier, there are three daemons operating in HDFS called NameNode, Secondary NameNode and DataNode. The HDFS cluster has single NameNode that runs on master node and maintains all the information about files and directories stored on the system. Moreover, there are several DataNodes, usually one per node within the cluster, which stores the blocks of data. The NameNode will provide the information to the client that which block resides on each Data Node. In other words, Data Nodes store and retrieve blocks upon instructions given from Name Node, and they report back to the NameNode at specified intervals with lists of blocks that they are storing. The Secondary NameNode merges the fsimage and the edits log files. fsimage is the snapshot of the file when NameNode started and edit logs are the sequence of changes made to the file system after NameNode started.

3.2.3 MapReduce

MapReduce is a parallel programming framework, considered as a distributed computation engine to process the data stored in file systems. Job Tracker and Task Tracker are two daemons that operate on MapReduce. When client submits a job to the JobTracker, it determines the location of data by communicating with NameNode. Then, it assigns the tasks to the available slave nodes through submitting it to its corresponding TaskTracker for execution. Moreover, JobTracker monitors all the tasks as they are running, if any task fails, it will resubmit the task possibly to a different node. On the other side, TaskTracker constantly communicate with JobTracker to signal the progress of the task in execution, through sending a heartbeat to JobTracker. If JobTracker fails to receive a heartbeat message from a TaskTracker within a specified amount of time, it will assume that the TaskTracker has become unresponsive and will resubmit the corresponding tasks to other nodes available in a cluster. In Hadoop framework programmers write their applications in MapReduce programming language, which has two functions called, Map and Reduce. These two functions are based upon key-value pairs. The input is passed to Map function, which generates a set of intermediate key-value pairs. The output of Map function is passed to the Reduce function, then all intermediate values are grouped together with the same key to produce the desired output (Dean and Ghemawat, 2008).

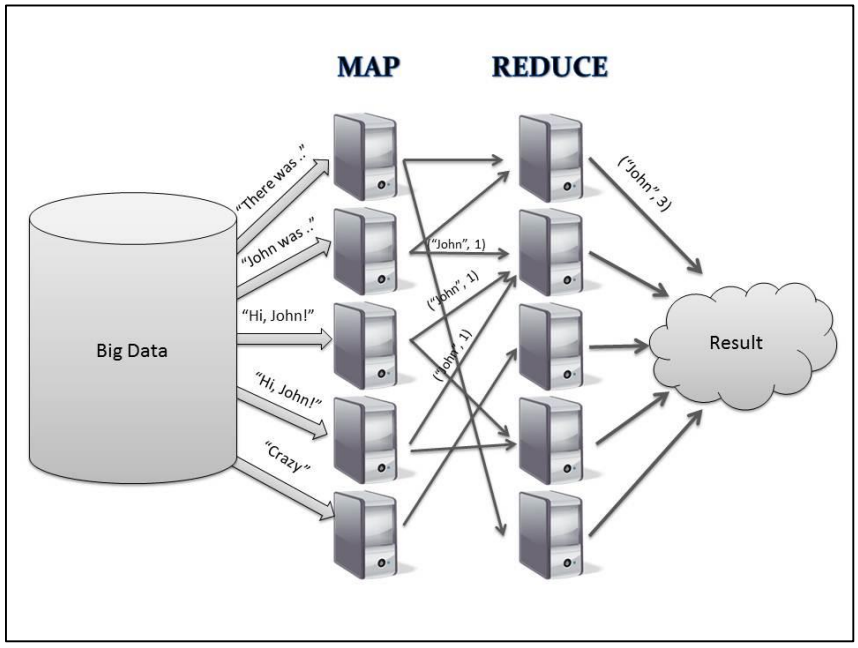

Figure 4: MapReduce Framework

3.3 Data Mining

Data mining has become a powerful technology to efficiently discover interesting information hidden among the huge amounts of data. Basically, it turns a large collection of data into useful knowledge. It is highly used in applications like Data analysis, Big data, Web mining, Business intelligence an Predictive analytics. Moreover, there are various methods for data mining, such as, Association rules mining, Cluster analysis, Regression, Text mining, Outlier detection and many more. In this research, clustering techniques will be used.

3.3.1 Clustering

One of the most significant method of data mining is clustering. It is an unsupervised learning technique, where the class label of input data is unknown. Clustering groups the data objects into clusters in such a way that the objects within a cluster are similar to each other and dissimilar to the objects in other clusters. Generally, there are two popular clustering techniques which are highly used for document clustering, namely; K-means and Agglomerative hierarchical clustering.

3.3.2 K-Means

K-Means is a well-known algorithm used for clustering (Kumar, Kiran and Prathap, 2013). K-means algorithm is a partitioning method that partitions a set of target objects into k clusters based on the similarity measure considered. The similarity measures in k-means assists us to find the degree of closeness or separation of target objects; therefore, clusters with high intra-cluster and low inter-cluster can be achieved. In document clustering, in order to achieve k number of clusters, k-means algorithm selects k documents from the corpus as initial centroids. Then it measures the similarity of every document to the initial centroids and assigns each document to the cluster which its value of similarity measure is minimum. In the next iteration, it calculates the new centroid of every cluster and reassign the documents based on the minimum similarity measure between the documents and the new centroids. This iterative process is continued until there is no change in clusters.

3.3.3 Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is a bottom-up strategy. It considers each object as an individual cluster and then iteratively merges the pair of clusters, which are more similar to each other, until all the objects are clustered into a single cluster. The results of the hierarchical clustering is displayed using Dendogram, which is like a tree diagram. (Steinbach, Karypis and Kumar, 2000).

3.4 Document Representation

Documents can be represented using Vector Space Model (VSM). In VSM documents are converted to vector of numbers based on the frequency of words, which means the number of times that particular word appears in the document. Assume D = {d_1, d_2, … ,d_n} is a data set that has m number of documents and T = {t_1, t_2,....., t_n} is a set of n distinct terms, which occurs in dataset D. Then the vector representation of document D is defined as,

$$t_d = \{tf(t_1, d), \ldots\ldots, tf(t_n, d)\}$$

where $tf(t, d)$ denote the frequency of term $t \in T$ in document $d \in D$.

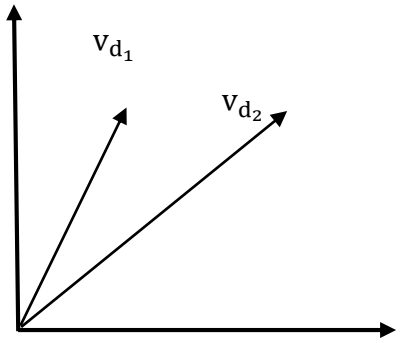

Figure 5: Documents Representation on VSM

As shown in Figure 5, the vector representation of two documents d_1 and d_2 is as follow:

$$v_{d_1} = \{tf(t_1, d_1), tf(t_2, d_1), \ldots\ldots, tf(t_n, d_1)\}$$

$$v_{d_2} = \{tf(t_1, d_2), tf(t_2, d_2), \ldots\ldots, tf(t_n, d_2)\}$$

Where $tf(t_n, d_1)$ denote the frequency of term $t \in T$ in document d_1 and $tf(t_n, d_2)$ denote the frequency of term $t \in T$ in document d_2.

3.5. Similarity Measures

Similarity measurement between documents is an elemental concept in text mining and information retrieval, which is widely used in many applications such as information retrieval, duplicate detection, automatic scoring, topic detection, document summarization, document clustering, and text classification (Gomaa and Fahmy, 2013). The main goal of similarity measurement is to quantify the degree of closeness or separation of target objects. In other words, similarity measure is a function that calculates the degree of similarity between the pair of documents. All the similarity measures map to the range of [-1, 1] or [0, 1]. The 0 or -1 represents minimal similarity, and 1 represents absolute similarity. Prior to clustering the documents, an appropriate similarity measure has to be determined, which is an important step in clustering the documents. Therefore, understanding the effectiveness of various similarity measures helps to choose the best one in order to achieve valid clustering output. Presently, there are variety of similarity measures available in literature. In this section, some most popular similarity measures used in detecting the similarity of documents is discussed.

Euclidean Distance: This is the most popular distance measure. It computes the distance between two vectors u and v, by calculating the square root of the sum of squared differences between corresponding elements of two vectors. The Euclidean distance of vector A and B is defined as,

$$\boldsymbol{d} = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \boldsymbol{y_i})^{\boldsymbol{2}}}$$

Cosine Similarity: Cosine similarity is an angle based measurement. It calculates the cosine of the angle between two vectors and helps us to find out how related to each other. The cosine similarity of vector A and B is defined as,

$$Cos(\theta) = \frac{A.B}{\|A\| \, \|B\|}$$

The cosine value is bounded between [0, 1]. If documents are similar, their vectors will be in the same direction from origin, thus, they form a relatively small angle, which its cosine value will be near to one. On the other hand, when two vectors are in different direction from origin, they form a wide angle and the value of the cosine is near to zero, consequently, the documents are dissimilar, and they map no similarity (Huang, 2008).

Dice's Coefficient: Dice Coefficient is calculated through multiplying two into the number of common terms in the compared documents divided by the total number of terms in both documents. Dice coefficient value map to ranges from [0, 1], where 0 represents non-overlapping and 1 represents a perfect agreement.

$$\text{Dice Coefficient} = \frac{2|A \cap B|}{|A| + |B|}$$

Jaccard Similarity: Jaccard coefficient or Tanimoto coefficient evaluates the similarity and non-similarity of two documents based upon the presence or absence of terms in documents. Ideally, it is calculated by dividing the total number of common terms between two documents by the entire number of terms that exists in at least one of the two documents. Moreover, the degree of similarity is a value between 0 and 1. If two documents are same and identical, the value is near to 1, conversely when it is 0 means that two documents are completely dissimilar.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Pearson Correlation Coefficient: This measure which provides how well two documents are related. Pearson correlation provides a linear relationship between two sets of data.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\,n\sum x^2 - (\sum x)^2]\ [n\sum y^2 - (\sum y)^2]}}$$

Latent Semantic Analysis (LSA): LSA provides a method to determine similarity of meaning of words by analyzing corpora. In LSA based on text, a matrix is constructed, each row of it represents a unique word, and each column represents a paragraph. The value

in each cell is the frequency of word that appears in the paragraph. Unfortunately, the constructed matrix tends to be very large and sparse; therefore, a mathematical technique called singular value decomposition (SVD) is used. SVD provides a reconstructed form of the matrix while preserving the least possible information. Finally, the words are compared by calculating the angle between the two vectors formed by any two rows (Thada and Jaglan, 2013) (Gomaa and Fahmy, 2013).

4. CHALLENGES

We have entered into an era, where tremendous amount of data is produced through multiple sources. As shown in Table 1 provided by Mckinsey Global Institute, which illustrates the different type of data generated and stored for various sectors. It's clear that the majority of data that is produced in different sectors is in the text format. Consequently, there is a requirement of analyzing this kind of data to extract the hidden knowledge and insights from them.

Table 1: Variety of Data across Sectors

	Video	**Image**	**Audio**	**Text/Numbers**
Banking	Medium	Medium	Low	High
Manufacturing	Medium	Medium	Low	High
Retail	Medium	Low	Low	High
Healthcare	Low	High	Low	High
Communications	High	Medium	High	High
Construction	Low	Medium	Low	Medium
Government	High	Medium	High	High
Education	High	Medium	High	Medium

The existing data mining techniques don't have the capability to capture, process and manage these heterogeneous, fast and high-volume data, so there is a requirement to use innovative techniques to solve such problems. The author in this research, will focus on using clustering techniques, which is a popular data mining technique. The input of the clustering algorithm is the corpus of documents, which are in the text format, because

analyzing text data is a very challenging task. Before clustering, it is important to choose a similarity measure which provides us coherent clusters. However, it is not clear that which similarity measure is best in document clustering. Therefore, it is highly significant to investigate the various similarity measures to select the best one. In addition, the ontology similarity measures are also effective in finding the similarity between documents due to the reason that they consider the sematic relation between the documents as well. Moreover, developing algorithms in MapReduce framework will help us in saving the execution time and space, which are upcoming challenges in this century.

5. PROBLEM DESCRIPTION

In document clustering, selecting an appropriate similarity measure to find the closeness between the documents is highly significant. To select that appropriate similarity measure is a challenging task, due to the reason that every set of documents shows a different result while applying the clustering algorithm on them. Moreover, the huge number of documents makes the process of finding similarity burdensome; therefore, there is a need to use a platform which will aid in paralleling, in order to reduce the time of computation. The best solution is using the leading framework Hadoop. In this research, different similarity measures will be investigated to find out the measures which provides the best solution after using the clustering algorithms. Finally, a new similarity measure will be proposed for designing and developing an efficient algorithm for clustering of text documents.

6. OBJECTIVES OF RESEARCH

The objectives of this research are as follows:

1. To survey the literature related to the research work done for investigating various similarity measures that are more useful in clustering the text documents.
2. To understand the effectiveness of various similarity measure to choose the best similarity measure in other to get the best clusters of text documents.
3. To evaluate the different similarity measures used for text documents using various tools.
4. To propose a new similarity measure to design and develop an efficient algorithm for text document clustering.

5. To implement the proposed algorithm and evaluating the results of achieved clusters through using clustering evaluation measures such as purity and entropy.

7. METHODOLOGY

The methodology to be adapted is divided into multiple phases as bellow:

Phase 1: Literature Review

In this phase the literature related to research work that has been done in the area of similarity measures for document clustering are reviewed.

Phase 2: Evaluation of Various Similarity Measures

The most useful similarity measures in document clustering is investigated. Moreover, there will be an investigation on different clustering techniques to find out which clustering techniques provides a better behavior in clustering the documents.

Phase 3: Design of New Similarity Measure

Based on the outcome and evaluations done in phase one and two, a new similarity will be designed and developed.

Phase 4: Developing an Efficient Document Clustering Technique

An efficient document clustering technique will be designed and developed based on the proposed similarity measure.

8. EXPECTED OUTCOME OF THE RESEARCH

A new similarity measure will be proposed, to design and develop an efficient document clustering technique, which can help organizations to group their documents to achieve an efficient browsing and searching.

9. PROPOSED TIME FRAME

Time / Work	1st Year		2nd Year		3rd Year		4th Year	
	Six months	Six months	Six months	Six months	Six months	Six months	Six months	Six months
Course work								
Literature and Review								
Publishing review papers								
Analyzing the existing works								
Designing and developing the new algorithms								
Paper publication of proposed work								
Final Report Writing and Editing								
Submission								

REFERENCES

Alam, A. and Ahmed, J., 2014, March. Hadoop Architecture and its issues. In Computational Science and Computational Intelligence (CSCI), 2014 International Conference on (Vol. 2, pp. 288-291). IEEE.

Apache hive™ (2011). Available at: https://hive.apache.org/ [Accessed: July 2016].

Apache HBase – Apache HBase™ Home (2007). [online] Apache HBase – Apache HBase™ Home. Available at: https://hbase.apache.org/ [Accessed July 2016].

Apache Oozie Workflow Scheduler for Hadoop (2011). Available at: http://oozie.apache.org/ [Accessed: July 2016].

Apache ZooKeeper (2010). Available at: https://zookeeper.apache.org/ [Accessed: July 2016].

Apache Mahout: Scalable machine learning and data mining. (2014), Available at: http://mahout.apache.org/ [Accessed: July 2016].

Apache Nutch™ (2004). Available at: http://nutch.apache.org/ [Accessed: Aug, 2016].

Apache Spark™ - lightning-fast cluster computing. Available at: http://spark.apache.org/ [Accessed: July 2016].

Apache storm (2016). Available at: http://storm.apache.org/ [Accessed: July 2016].

Breakthrough Analysis. (2008). Unstructured Data and the 80 Percent Rule. Available at: https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/ [Accessed: Aug 2016].

Bruza, P. and Song, D., 2001. Discovering Information Flow using a High Dimensional Conceptual Space. In Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval.

Dean, J. and Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), pp.107-113.

El-Hamdouchi, A. (1989). Comparison of Hierarchic Agglomerative Clustering Methods for Document Retrieval. The Computer Journal, 32(3), pp.220-227.

Elsayed, T., Lij, J. and W., Oard, D. (2008). Pairwise Docuent Similarity in Large Collections with MapReduce. In: HLT-Short 08 Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Technologies, pp. 265-268.

Gomaa, W.H. and Fahmy, A.A., 2013. A survey of text similarity approaches. International Journal of Computer Applications, 68(13).

Ghemawat, S., Gobioff, H. and Leung, S.-T. (2003) 'The Google file system', ACM SIGOPS Operating Systems Review, 37(5), p. 29.

Hirschberg, D.S., 1977. Algorithms for the longest common subsequence problem. Journal of the ACM (JACM), 24(4), pp.664-675.

Huang, A., 2008, April. Similarity measures for text document clustering. In Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, pp. 49-56.

Ishwarappa and Anuradha, J. (2015) 'A brief introduction on big data 5Vs characteristics and Hadoop technology', Procedia Computer Science, 48, pp. 319–324.

Kumar, A., Kiran, M. and Prathap, B.R., 2013, July. Verification and validation of MapReduce program model for parallel k-means algorithm on Hadoop cluster. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on (pp. 1-8). IEEE.

Niyigena, P., Zuping, Z., Li, W. and Long, J. (2015). Efficient Pairwise Document Similarity Computation in Big Datasets. IJDTA, 8(4), pp.59-70.

Steinbach, M., Karypis, G. and Kumar, V., 2000, August. A comparison of document clustering techniques. In KDD workshop on text mining (Vol. 400, No. 1, pp. 525-526).

Shvachko, K., Kuang, H., Radia, S. and Chansler, R., 2010, May. The hadoop distributed file system. In 2010 IEEE 26th symposium on mass storage systems and technologies (MSST) (pp. 1-10). IEEE.

Thada, V. and Jaglan, D.V., 2013. Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. International Journal of Innovations in Engineering and Technology (IJIET), 2(4), p.202.

Welcome to Apache pig! (2007). Available at: https://pig.apache.org/ [Accessed: July 2016].

Welcome to Apache Flume - Apache Flume (2009). Available at: https://flume.apache.org/ [Accessed: July 2016].

Wu, X., Zhu, X., Wu, G.Q. and Ding, W., 2014. Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), pp.97-107.

White, T., 2012. Hadoop: The definitive guide. “O’Reilly Media, Inc.".